Experience in Ontology Engineering for a Multi-Agents Corporate Memory System

Fabien Gandon
ACACIA project - INRIA, 2004, route des Lucioles, B.P. 93
06902 Sophia Antipolis, France - Fabien.Gandon@sophia.inria.fr

Abstract

XML and multi-agents technologies offer a number of assets for corporate memory management. Since ontologies appear as a key asset in the new generation of information systems and also in the communication layer of multi-agents systems, it comes with no surprise that it stands out as a keystone of multi-agents information systems. Here, we briefly describe our approach and motivations and then focus on the first elements of our return on experience in building an ontology for such a system.

1 Introduction

In the last decade information systems became backbones of organizations and the industrial interest in methodologies and tools enabling capitalization and management of corporate knowledge grew stronger. A corporate memory is an explicit, disembodied and persistent representation of knowledge and information in an organization, in order to facilitate their access and reuse by members of the organization, for their tasks [Rabarijaona *et al.*, 2000]. The stake in building a corporate memory management system is the coherent integration of this dispersed knowledge in a corporation with the objective to "promote knowledge growth, promote knowledge communication and in general preserve knowledge within an organization" [Steels, 1993]. ACACIA, our research team, is part of the CoMMA project (IST-1999-12217) funded by the European Commission, aiming at implementing a corporate memory management framework based on several emerging technologies: agents, ontologies, XML, information retrieval and machine learning techniques [CoMMA, 2000]. These technical choices are mainly motivated by three observations. **(1) The memory is, by nature, an heterogeneous and distributed information landscape.** The corporate memories are now facing the same problem of precision and recall than the Web. The initiative of a semantic Web is a promising approach where the semantics of documents is made explicit through metadata and annotations to guide the later exploitation of these documents. XML enables us to build a structure around the data, and RDF (Resource Description Framework) allows resources to be semantically annotated. **(2) The tasks as a whole to be performed on the memory are, by nature, distributed and heterogeneous.** So we envisaged a distributed and heterogeneous system to explore and exploit this information landscape: a multi-agents system (MAS). It allows the resources to remain localized and heterogeneous while enabling to capitalize an integrated and global view of the memory thanks to cooperating software agents distributed over the network and having different skills and roles to support the memory tasks. The heterogeneity and distribution of the MAS is an answer to the heterogeneity and the distribution of the corporate memory. **(3) The population of the users of the memory is, by nature, heterogeneous and distributed in the corporation.** Agents will also be in charge of interfacing users with the system. Adaptation and customization are a keystone here and we are working on machine learning techniques in order to make agents adaptive to the users and the context. This goes from basic customization to user's habits and preferences learning, up to push technologies based on interest groups and collaborative filtering.

2 Approach Overview

Compared to the Web, a corporate memory has more delimited and defined context, infrastructure and scope ; the existence of a community of stakeholders means that an ontological commitment is conceivable to a certain extend. So far, the enterprise modeling field has been mainly concerned with simulation and optimization of the design of the corporate production system but last decade changes led enterprises to become aware of the value of their memory and the fact that enterprise models have a role to play in this application too. The corporation has its own organization and infrastructure ; this state of affair can be formally made explicit to guide the corporate memory activities involved, for instance, in the new employee integration and the technology monitoring scenarios of CoMMA. This enables the system to get insight into the organizational context and environment and to intelligently exploit it in interactions between agents and between agents and users. Likewise, the users' profile captures all aspects of the user that were identified as relevant for the system behavior. It contains administrative information and directly explicited

preferences that go from interface customization to topic interests. It also positions the user in the organization: role, location and potential acquaintance network. In addition to explicitly stated information, the system will derive information from the usage made by the user. It will collect the history of visited documents and possible feedback from the user, as well as the user's recurrent queries, failed queries, and from this it can learn some of the user's habits and preferences. These derived criterions can then be used for interface purposes or push technology. Finally the profiles enable to compare users, to cluster them based on similarities in their profiles and then use the similar profiles to make suggestions.

The figure 1 gives the OSA modeling architecture use in CoMMA. Our approach is : (1) to apply knowledge engineering techniques to provide the conceptual vocabulary needed by the scenarios and to formalize this ontology in RDF using the RDF Schema (2) to describe the organizational state of affair and users' profile in RDF statements (3) to structure the corporate memory with RDF annotations based on the ontology and referencing the state of affair (4) to use the annotations, the state of affair and the ontology through inferences in order to search, manage and navigate into the memory. As shown in figure 1, the ontology and the state of affair form the model ; the archive annotations will depend on both. The state of affair and the annotations are instances of the RDF schema : the ontology is at the intensional level whereas the state of affair and the annotations are at the extensional level. The ontology, the state of affair and the annotations are tightly linked and will evolve as a whole in a prototype life cycle style.

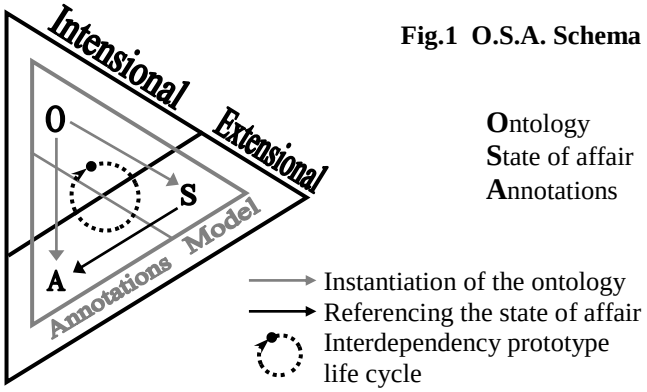

Fig.1 O.S.A. Schema

CoMMA is an heterogeneous Multi-Agents Information System (MAIS) supporting information distribution. The duality of the definition of the word 'distribution' reveals two important problems to be addressed : (a) Distribution means dispersion, that is the spatial property of being scattered about, over an area or a volume ; the problem here is to handle the naturally distributed data, information or knowledge of the organization. (b) Distribution also means the act of spreading or apportioning ; the problem then is to make the relevant pieces of information go to the concerned agent. In a MAS, distribution is handled through cooperation so in our case, agents must be able to communicate with the others to delegate tasks or solve queries. The content of the exchanged messages relies on the ontology. The agents play roles and are organized in societies as described in [Gandon *et al.*, 2000]. In order to manipulate the ontology, the annotations, and infer from them, the agents import modules from CORESE a prototype of a search engine enabling inferences on RDF annotations by translating the RDF triplets to Conceptual Graphs and vice versa [Corby et al., 2000].

3 Engineering an ontology

Following Caroll [1997] we used *scenarios* to capture end-users' needs in their context. They enable us to focus on the specific aspects of knowledge management involved in our case, to capture the whole picture and a concrete set of interaction sequences, and to view the system as a component of a knowledge management solution for a company. A scenario template was proposed, suggesting key aspects to be considered when describing a scenario and collecting data. This helps define the scope of our intervention and thus the scope of the ontology. Scenario analysis produced reports which are extremely rich story-telling documents and therefore good candidates to be included in the corpus of a terminological study.

Several techniques exist for *data collection*, we used three of them: semi-structured interview, observation and document analysis. Data collection also included the study of existing ontologies: the Enterprise Ontology [Uschold *et al.*, 1998], the TOVE Ontology [TOVE, 2000], the Upper Cyc Ontology [Cyc, 2000], the PME Ontology [Kassel *et al.*, 2000] and the CGKAT & WebKB Ontology [Martin and Eklund, 2000 ; Martin, 1996]. The reuse of ontologies is both seductive (saves time, efforts and favors standardization) and difficult (commitments and conceptualizations have to be aligned between the reused ontologies and the needed one). These ontologies have not been imported directly, the best way for us to use them was to start from their informal version in natural language. Natural language processing tools could help this analysis, and translators between formal languages could ease reuse. Reused sources have to be pruned ; scenarios capture the scope of the intervention and a shared vision of the stakeholders, they can be used to decide whether or not a concept is relevant e.g.: the 'ownership' relation of the Enterprise Ontology was not reused in our ontology because this relation does not seem exploitable in our scenarios. We also considered other informal sources: some very general ones helped us structure upper parts of some branches e.g.: the book '*Using Language*' from H.H. Clark inspired the branch on representation systems ; others very specific enabled us to save time on enumerating some leaves of the taxonomical tree e.g.: the MIME standard for electronic format description. The systematic use of dictionaries or available lexicons is good practice. In particular, the meta-dictionaries have proved to be extremely useful. They enable access to a lot of dictionaries and therefore one can easily compare definitions and identify or build the one that correspond to the notion one wants to introduce. We made extensive use of [OneLook, 2000].

The *candidate terms* were collected in a set of informal tables. The next step is to produce consensual definitions to build the concepts defined 'in intension'. At this point, labeling concepts with one term is both convenient and dangerous. It is a major source of 'ambiguity relapse' where people relapse in ambiguity using the label terms according to the definition they associate to it and not the definition actually associated to it during the semantic commitment. The explicit representation and the existence of management functionality for terminological aspects in tools assisting ontologists are real needs. The obtained concepts were organized in a taxonomy: we started regrouping concepts firstly in an intuitive way, then iteratively organizing and reviewing the structure. We studied several principles to build the taxonomical tree: the extended Aristotelian principles in [Bachimont, 2000], the semantic axis in [Kassel *et al.*, 2000], and the extensive work of Guarino and Welty [Guarino, 1992; Guarino and Welty, 2000]. The main problem is that, as far as we know, no tool is available to help an ontologist apply these principles easily and independently of a formalization language; it can become a titanic work to apply these theories to large ontologies.

The way to design an ontology is still debated in the knowledge engineering community. There is a tendency to distinguish between three approaches: *Bottom-Up, Top-Down and Middle-Out*. We are not convinced that there exists such a thing as a purely top-down, bottom-up or middle-out approach. They seem to be *three complementary perspectives of a complete methodology* with concurrent processes present and at work at different levels of depth (bottom, middle or top) and different detail grains (concepts or groups of concepts). We shall not deny that for a given case, an approach can mainly rely on one perspective, but we would not oppose them as different approaches: when engineering an ontology, an ontologist should have the tasks defined in these three perspectives on the go at one time. In our case, some tasks were performed in parallel in the different perspectives, e.g. : we studied existing top-ontologies and upper parts of relevant ontologies to structure our top part and reuse parts of existing taxonomies (top-down approach); we studied different branches, domains, micro-theories of existing ontologies as well as core subjects identified during data collection to understand what were the main areas we needed and group candidate terms (middle-out approach); we exploited reports from scenario analysis and data collection traces to list scenario specific concepts and then started to regroup them by generalization (bottom-up approach). The different buds (top concepts, core concepts, specific concepts) opening out in the different perspectives are the origins of partial sub-taxonomies. The objective then is to ensure the joint of the different approaches and an event in one perspective triggers checks and tasks in others.

This approach resulted in a more or less *three-layered ontology*: (1) A very general top (2) A very large middle layer divided in two main branches: one generic for corporate memory domain and one dedicated to the topics of the application domain (3) An extension layer which tends to be scenario and company specific with internal complex concepts. We obtained three semi-informal tables (concepts, relations and attributes) with the following columns: (1) the label of the concepts / relations / attributes, (2) the concepts linked by the relation or the concept and the basic type linked by the attribute, (3) the closet core concept or the thematic fields linked by the relation, (4) the inheritance links, (5) synonymous terms for the label, (6) a natural language definition to try to capture the intension, and (7) the collection source. This last column introduces the principle of traceability and it is interesting for the purpose of abstracting a methodology from the work done. It enables to know what sort of contribution influenced a given part of the ontology and to trace the effectiveness of reuse. However this is by far not enough and the complete design rationale of the ontology should be captured in order to help people understand and may be commit to or adapt it.

The final formal degree of the ontology depends on its intended use. *The goal of the formalization task is not to take an informal ontology and translate it into a rigorously formal ontology, but to develop the formal counterpart of interesting and relevant semantic aspects of the informal ontology* in order to obtain a documented (informal description possibly augmented by navigation capabilities from the formal description) operational ontology (formal description of the relevant semantic attributes needed for the envisioned system). The formal form of an ontology must include the natural language definitions, comments, remarks, that will be exploited by humans trying to appropriate the ontology. This also plays an important role for documenting the ontology and therefore for ontology reuse, reengineering and reverse-engineering.

In our case, the last step of formalization was the *translation of semi-informal tables in RDF*. Thanks to the XML technology we managed to keep the informal view through XLST style sheets: (a) a style sheet recreates the table of concepts (b) a second one recreates the table of relations and attributes (c) a last one proposes a new view as a tree of concepts with their attached definition as a popup window following the mouse pointer. This pop-up is a first attempt to investigate how to proactively disambiguate navigation or querying: before the user clicks on a concept, the system displays the natural language definition inviting the user to check his personal definition upon the definition used by the system so as to avoid misunderstandings. The second interesting point of that view is that if the user clicks on a concept he obtains all the instances of this concept and its sub-concepts, so this view is a link between the intensional level and the extensional one.

The design of an ontology is an iterative maturation process, it follows a prototype life-cycle [Fernandez *et al.*, 1997]. As an example, one of the problems spotted when reviewing the ontology was the redundancy ; for instance we found that annotating a document as multi-modal is redundant with the fact that we annotated it with the different modes it uses. So we decided that the multi-modal was not a basic annotation concept and that it should be a defined concept derived from other existing concepts where

possible. However the notion of defined concept, does not exist in RDFS, and we will have to extend the schema as proposed in [Delteil *et al.*, 2001].

The first draft of the ontology was a good step for feasibility study and first prototypes, but it comes with no surprise that the prototype life-cycle is time consuming. Moreover the ontology is a living object the maintenance of which has consequences beyond its own life-cycle : what happens to the annotations written thanks to this conceptual vocabulary when a change occurs in the ontology? Deletion and modification obviously raise the crucial problem of coherence and correction of the annotation base. But an apparently innocuous addition of a concept also raises the question of the annotations using a parent concept of the new concept and that could have been more precise if the concept had existed when they were formulated: should we review them or not ? These problems are obviously even more complex in the context of a distributed system. Finally, an ergonomic representation interface is a critical factor for the adoption of the ontology by the users; if the user is overloaded with details or lost in the meanderings of the taxonomy he will never use the system and the life-cycle of the ontology will never complete a loop. We are investigating that point, and the terminological level seems very important here too.

4 Conclusion

Ontologies are a keystone of multi-agent systems and play an important role in the new generation of information systems, therefore they will clearly become a central component of MAIS and they surely do in CoMMA. Our experience gave rise to several expectations and to be able to manage, share and discuss the growing ontology, we would definitively need an integrated environment with: (a) improved interfaces for representation, navigation and manipulation of ontologies (b) natural language processing tools to semi-automate the analysis of the extensive part of the resources that are textual (c) facilities for applying the results from theoretical foundations of Ontology and help ontologists check their ontologies (d) tools to manage the versioning of the ontology and all that has been built upon it (annotations, models, inferences...) and to capture the design rationale. Finally work is needed to help make explicit and preserve the intensional semantic structure of the computational level. If the new generation of AI agents is to be based on an explicit conceptualization, this must not be limited to the knowledge exchanged currently, it must include the action performed on it with both their intension and intention.

Acknowledgements

I warmly thank my colleagues Rose Dieng, Olivier Corby and Alain Giboin, the CoMMA consortium and the European Commission funding the CoMMA project.

References

[Bachimont, 2000] Bachimont, Engagement sémantique et engagement ontologique: conception et réalisation d'ontologies en ingénierie des connaissances, In J. Charlet et al., *Ingénierie des connaissances Evolutions récentes et nouveaux défis*, Eyrolles

[Caroll, 1997] Caroll, Scenario-Based Design, In Helander et al. *Handbook of Human-Computer Interaction.*, Chap. 17, Elsevier Science B.V.

[CoMMA, 2000] CoMMA, Corporate Memory Management through Agents, In *Proc. E-Work E-Business*

[Corby *et al.*, 2000] Corby, Dieng, Hébert. A Conceptual Graph Model for W3C Resource Description Framework. In *Proc. ICCS'2000*

[Cyc, 2000] www.cyc.com/cyc-2-1/cover.html

[Delteil *et al.*, 2001] Delteil, Faron, Dieng, Extension of RDFS based on the CG formalism, In *Proc. ICCS'01*

[Fernandez *et al.*, 1997] Fernandez, Gomez-Perez, Juristo. METHONTOLOGY: From Ontological Arts Towards Ontological Engineering. In *Proc. AAAI97 Symposium Ontological Engineering*

[Gandon *et al.*, 2000] Gandon, Dieng, Corby, Giboin, A Multi-Agents System to Support Exploiting an XML-based Corporate Memory, In *Proc. PAKM'00*

[Guarino and Welty, 2000] Guarino, Welty, Towards a methodology for ontology-based model engineering. In *Proc. ECOOP-2000 Workshop Model Engineering*.

[Guarino, 1992] Guarino, Concepts, Attributes, and Arbitrary Relations: Some Linguistic and Ontological Criteria for Structuring Knowledge Bases. In *Data and Knowledge Engineering* 8

[Kassel *et al.*, 2000] Kassel, Abel, Barry, Boulitreau, Irastorza, Perpette, Construction et exploitation d'une ontologie pour la gestion des connaissances d'une équipe de recherche. In *Proc. IC'00*

[Martin and Eklund, 2000] Martin, Eklund, Knowledge Indexation and Retrieval and the Word Wide Web. *IEEE Intelligent Systems special issue Knowledge Management over the Internet*.

[Martin, 1996] Martin, Ph.D. Thesis Exploitation de Graphes Conceptuels et de documents structurés et Hypertextes pour l'acquisition de connaissances et la recherche d'informations, University of Nice Sophia Antipolis

[OneLook, 2000] www.onelook.com/

[Rabarijaona *et al.*, 2000] Rabarijaona, Dieng, Corby, Ouaddari, Building and searching a XML-based Corporate Memory, *IEEE Intelligent Systems Special Issue Knowledge Management and Internet* 56-64

[Steels, 1993] Steels, Corporate Knowledge Management. In Barthès ed., *Proc. ISMICK'93*

[TOVE, 2000] www.eil.utoronto.ca/tove/ontoTOC.html

[Uschold *et al.*, 1998] Uschold, King, Moralee, Zorgios, The Enterprise Ontology. In Uschold and Tate *The Knowledge Engineering Review Special Issue on Putting Ontologies to Use*, Vol. 13